Deep Learning - 3. Assignment - LSTM's & Transformers

Nejc Ločičnik (63180183)

1 Introduction

This assignment is about the LSTM and Transformer like architectures for text generation. We evaluate text outputs for different sampling methods and sequence lengths used during training. Additionally both architectures are tested with varied amount of layers (amount of LSTM cells and Transformer blocks) and training parameters.

2 LSTM

The task was to implement a LSTM Cell and implement it into a simple LSTM model. I Additionally added a multilayered LSTM model.

2.1 Multilayered LSTM

1, 2, 3 and 5-layered LSTM was tested. The best results were obtained with 2 layers.

```
Top-K sampling -----------------
O Romeo, wherefore art thou art, wash myst death
Will burn my fatitions; these stays, which will hear him
To before them have setting to the cause.

LUCENT:
When yet, as you have but as they have so meat
The seats and soldier, but here but belonds
My surposed frient, the peace is not to my father's
cold its of chambet, would
-----------------------------------------------
Greedy sampling ---------------
O Romeo, wherefore art thou art a thousand death
To see the seat of the present to the people,
The proud and the present to the proud and soul
To the present to the proud and soul
The state the present to the proud and soul
That we are the present to the proud and soul
Is the state that he was the state and here,
The proud an
```

Figure 1: Output of 2-layered LSTM with Top-K and Greedy sampling.

As we can see in figure 1, the results depend quite a lot on the used sampling method. The Top-K sampling method samples the 5 ($k = 5$) most probable tokens using the multinomial distribution, while Greedy sampling always takes the most probable token. Top-K therefore provides more varied/random outputs, while Greedy almost always gets caught in exposure bias (mode collapse), where it starts repeating the same thing over and over. The length of the repeated part depends a lot on the sequence length during training and partly on the models architecture.

I said the 2-layered LSTM performed the best, because it was the only model setup that still produced some variation with Greedy sampling (even though its still the same words shifted around).

I also tried changing the training parameters without much success. Changing learning rate in any way seemed to have negative effects on the training, the default value of 0.005 worked best. Increasing hidden state size seemed to improve the training (looking purely at the final loss), but the difference in the text output was not noticeable.

2.2 Sequence length impact

The length of the training sequences determines how far back is the model trained to retain relation/dependencies of the words. Short training sequence lengths will focus the training on more local properties, like single sentence structure, while longer sequence lengths can take into account more global properties like paragraph structure. Obviously more local properties are still contained in longer sequences so having a bigger sequence length should be better (this becomes a memory problem though).

The result of different sequence lengths is most noticeable in the Greedy sampling as the repeated sequence will be shorter (figure 2 - above) or longer (figure 2 - below).

```
Greedy sampling ----- chunk_len = 64 ------
O Romeo, wherefore art thou art thou art the sun
The sun the sun the sun the sun the sun
The sun the sun the sun the sun the sun

Greedy sampling ----- chunk_len = 256 ------
O Romeo, wherefore art thou shalt see the seas,
And therefore he shall be so many things and so,
That I will be so much a thousand for thee,
And therefore he shall be so many things and so,
That I will be so much a thousand for thee,
```

Figure 2: Output of Greedy sampling with different sequence lengths used during training.

3 Transformer-like network

The requirement for the Transformer-like network was to implement the attention mechanism of the model. The attention mechanism consists of two parts.

The first part is the Scaled dot-product, the actual calculation of attention. This operation enables the model to focus more effectively on relevant information in the input sequence.

The second part is the Multi-head attention, which additionally enhances the expressiveness of the Scaled dot-product by allowing the model to attend to different aspects of the input sequence simultaneously. It achieves this by performing multiple attention operations in parallel, each with different weights.

This allows the model to capture diverse relationships and dependencies in the input sequence. For example one head might learn simple sentence structure, while another might learn how to chain sentences.

3.1 Masking attention

The reason for the masking operation in the attention mechanism (Scaled dot-product) is to enforce the autoregressive property (enables generation of new tokens) while training. During the training, we want the model to rely purely on the information provided in the current and previous positions to predict the next token. By masking future positions we ensure the model can only attend to the current and preceding positions in the input sequence, which prevents information leakages from future positions.

The masking operation is usually done by adding large negative values ($-inf$) to the future positions in the attention matrix. This ensures that probabilities corresponding to these positions become effectively zero after the softmax operation, preventing the model from attending to those positions.

3.2 Transformer results

I attempted to fine-tune some training parameters, but everything (schedule, decay, learning rate change) seemed to make the training not converge or do so very slowly, so I made some changes to the model architecture instead. I increased the number of Transformer blocks to 8, increased number of heads to 10 and number of neurons per head to 80. I didn't try other types of regularisation as the model already includes layer normalisation. The output is shown in figure 3.

The difference between Top-K and Greedy sampling seems to be text coherence. Top-K paragraphs seem unrelated to one another, everything seems too random. Greedy sampling sentence structure seems even worse, but the paragraphs seem like a talk between 3 people. The whole text as a whole seems more structured and less random with Greedy sampling. Other than that I can't see a clear difference.

3.3 Sequence length impact

Because of memory issues for longer sequence lengths the Transformer model for this part used the default parameters.

The results were a bit surprising, I thought they will be similar to LSTM's. The main difference in increasing/decreasing sequence length during training seems to be the format of the text. Lower sequence length produced more dense text - monologues (see figure 4), compared to longer sequence length which produced very sparse text - conversations (see figure 5). I won't comment on sentence coherence and so on as it all seems similar to me.

```
Top-K Sampling:
----------------------------------------
Here's to my love! O true apothecary! Thy drugs are quick.
Is this the king and hange of mine, I love,
That I myself, but stoood as I recean,
The mark'd means o' the father speak:
Yet speak before, sir, speak not nor son.

DUCHESS OF YORK:
I had a bed and and nail--shearted by the king,
And that my means to blind that thou bidst good,
Lest thou noble thing, but that I cann'd the give.

HORTENSIO:
Yes, my good gentle sir; I lesseigh you
How many did:dow b

Greedy Sampling
----------------------------------------
Here's to my love! O true apothecary! Thy drugs are quick.

LADY CAPULET:
And then, I'll give me thy hand, that live,
So stillengess drew shoutes, that did beauty,
And, tend times-to night to-day, be gone.

GLOUCESTER:
Go, and one too, that folly conjuration and in
The sallt needle by the marriage.

GLOUCESTER:
Indeeed, I cannot get your honour.

LADY ANNNE:
If I I would trade:
I knew they can not; but I'lll give me leave,
And so she cannible number with
```

Figure 3: Output of Transformer-like model with Top-K and Greedy sampling.

4 Conclusion

In conclusion, both LSTM and Transformer architectures showed promising results for text generation tasks. However, evaluating and comparing generated text outputs remains challenging due to the subjective nature of text quality assessment. Factors such as coherence, relevance, and fluency depend a lot on the context and intended use of the generated text, making it difficult to define clear evaluation metrics. Despite these challenges, both architectures offer valuable insights into the capabilities and limitations of deep learning models for text generation tasks.

```
Top-K Sampling:
----------------------------------------
Here's to my love! O true apothecary! Thy drugs are quick.

GLOUCESTER:
I do believe this wretch charrged willl death.


LUCIO:
These greatsorse satisfy to his person, and himself
How sailty too beast all posssed; when he stands he
gentle haste, with a winged bloow the cause of this place,
Which, wert thou come
Wate home: though I have sperit
In posssible ass stem, intrussed me
IntreaN, and alll the world, thou wilt plead,
To fear off the starm of hell a

Greedy Sampling
----------------------------------------
Here's to my love! O true apothecary! Thy drugs are quick.

KING EDWARD IV:
No, Greough: nor go fine hath as steep ass
Werst fear the pent to off an oath?
O my wife, thou seen me not? Well will are you?


Pedast:
Your worrrong me nor grant to you come;
For such ffair aunt or a place of the dead.
I wish me tome means, and I him, and welll
Condemned, to me to them and the placefe of the world:
When I shalll, be sadd, the sigester is alll.
The word of Gor h
```

Figure 4: Output when decreasing chunk length to 64 during training.

```
Top-K Sampling:
Here's to my love! O true apothecary! Thy drugs are quick.

First Gentleman:
Thy father is mortal married on my son, thy words
as more beholding as thee, thy sullen-ta'e took us.

GLOUCESTER:
The glory enough the man that would live,
Sun the tears of their sheeps each abides.

LARTIUS:
So, then.


CAMILLO:
I talke of the headgest what I was your heart,
When my old would have it yearson and take
The people of your breath your power youth
Where as fut as a

Greedy Sampling
Here's to my love! O true apothecary! Thy drugs are quick.

JULIET:
What's thy noise?

Nurse:
The servoused and bids us: by this toorgetorment,
crown thee for my hand is privyelges.

JULIET:
And so it,--as I told my langthemen,--I am the drop
In will flowere brows on the lord:---
Alack, for sleeep,--

LEONTES:
Where I thank your best your loves--onda---
Heave your past; catcher, and whither wille.

PETHVIRGONE:
Sirr, you have no cause.
```

Figure 5: Output when increasing chunk length to 256 during training.